1578–1584 *Nucleic Acids Research, 1999, Vol. 27, No. 7*

mRNAs have greater negative folding free energies than shuffled or codon choice randomized sequences

William Seffens[1,2,*] and David Digby[1]

[1]Department of Biological Sciences and [2]Center for Theoretical Study of Physical Systems, Clark Atlanta University, 223 James Brawley Drive, South West, Atlanta, GA 30134, USA

Received as resubmission February 3, 1999; Revised and Accepted February 12, 1999

ABSTRACT

An examination of 51 mRNA sequences in GenBank has revealed that calculated mRNA folding is more stable than expected by chance. Free energy minimization calculations of native mRNA sequences are more negative than randomized mRNA sequences with the same base composition and length. Randomization of the coding region of genes yields folding free energies of less negative magnitude than the original native mRNA sequence. Randomization of codon choice, while still preserving original base composition, also results in less stable mRNAs. This suggests that a bias in the selection of codons favors the potential formation of mRNA structures which contribute to folding stability.

INTRODUCTION

Control of gene expression is known to occur at any of the events from promotion of transcription to stabilization of the mature polypeptide product. Several studies have demonstrated that mRNA stability may be an important factor in gene expression for certain genes (1–3). Structural RNA features are suspected to be involved in the regulation of mRNA degradation in those cases. Several authors have suggested that the choice of codons in eukaryotic genes may be constrained by effects other than the frequency of codons in the whole gene (4–6). This is shown by the occurrence of specific codons at specific positions in a large set of genes with a frequency larger than expected by chance (5). Statistical non-randomness in the occurrence of codons was demonstrated for 96 eukaryotic genes, including the actin and β-globin gene families (7). For a message coding 100 amino acids, there are ~3^{100} or 10^{48} different combinations of bases using synonymous codons coding for the same polypeptide. This work tested the hypothesis that codon choice is biased to generate mRNAs with greater negative folding free energies.

MATERIALS AND METHODS

mRNA sequences were selected from GenBank using programs in the Wisconsin Group GCG software package (8). mRNA sequence files were randomly selected from GenBank with short Locus descriptors (limited to eight or nine characters) and which possessed sufficient information in the Features annotation to reconstruct the sequence of the mRNA. Fifty-one mRNA sequences were selected from the GenBank sequence database possessing the following properties identified in the Features annotation: (i) mRNA +1 start site identified; (ii) MET start codon identified; (iii) termination codon identified; (iv) poly-A site or signal identified; and (v) the mRNA sequence must be <1200 bases long. A variety of sequence files were examined from diverse species including prokaryotes, plants, invertebrates and higher animals (Table 1). These mRNA sequences were *in silico* folded using Zuker's MFOLD program from UWGCG using a VAXStation 4000 or SUN Ultra computer (9).

Each *in silico* folding free energy of a native mRNA was compared with folding free energies calculated from mRNA sequences randomized by one of six different procedures. In the first randomization procedure, each native mRNA sequence was randomized at least 10 times using the SHUFFLE program of UWGCG. SHUFFLE randomizes the order of bases in a sequence keeping the composition constant. These randomized sequences (termed 'whole-random' sequences) were folded and the free energies averaged. In the second randomization procedure, the native sequences were randomized only within the coding region, yielding 'CDS-random' sequences. These sequences contained unmodified 5′ and 3′ untranslated regions (UTRs). In the third randomization procedure, codons were shuffled within the coding sequence only, yielding 'codon-shuffled' sequences. These contain unmodified UTR sequences of the respective native mRNA, and code for a polypeptide with identical amino acid composition yet different amino acid sequence. A program (RNAshuffle) was written in FORTRAN using the GCG software library that randomized only the codon choice to produce 'codon-random' sequences for the fourth randomization procedure. Codon-random sequences have the same nucleotide base composition and translated polypeptide product as the respective native mRNA. The fifth randomization procedure was a modification of the previous codon-random algorithm without constraining the base composition. All codons were allowed to be equally likely. The resulting sequences also coded for the same polypeptide as the native mRNA, but the base composition was generally more G+C rich. These sequences were labeled as 'codon-flat'. The final randomization procedure tested the UTRs by shuffling both of the UTRs together while leaving the CDS unchanged. The lengths of the UTRs remained the same in this procedure called 'UTR-random'.

*To whom correspondence should be addressed at Biology Department, Clark Atlanta University, 223 James Brawley Drive, South West, Atlanta, GA 30134, USA. Tel: +1 404 880 6822; Fax: +1 404 880 6756; Email: wseffens@cau.edu

Table 1. Characteristics of selected mRNA sequences

Gene	Length	5′ UTR	Coding	3′ UTR	Comments
Invertebrate sequences					
DROCSKB	946	22	648	276	casein kinase II beta
DROMETO	301	69	132	100	metallothionein protein
DROSIST	782	66	570	146	sister-less-a protein
DROTU4A	625	62	507	56	vitelline membrane protein
DROUBXDR	663	115	306	242	ultrabithorax bithoraxoid
DROVMP	434	29	351	54	vitelline membrane protein
Bacterial sequences					
ECOADD	1039	31	999	9	adenosine deaminase
ECOALKA	887	19	849	19	DNA glycosylase
ECOCMA	901	59	816	26	colicin M activity peptide
ECODAPA	927	24	879	24	DHDP synthetase
Human sequences					
HUMALR	1132	60	978	94	aldehyde reductase
HUMCAL	791	70	426	295	calcitonin
HUMCALCI	681	39	426	216	calcitonin
HUMGRP5E	797	55	447	295	gastrin-releasing peptide
HUMGST	909	73	468	368	glutathione S-transferase
HUMHEMBP	822	48	669	105	eosinophil basic protein
HUMHIS4	390	28	312	50	histone H4
HUMHPBS	810	61	510	239	benzodiazepine receptor
HUMIFNAB	1041	31	570	440	interferon alpha-b
HUMIFNAC	963	46	570	347	interferon alpha-c
HUMIFNAF	985	8	570	407	interferon alpha-f
HUMIFNAH	985	56	570	359	interferon alpha-h
HUML12A	612	68	498	46	ribosomal L12 protein
HUMOGC	891	311	336	244	protein
HUMP1BX	480	35	243	202	secretory protein
Mouse sequences					
MMU03711	618	42	453	123	demilune cell-specific
MUSCASK	785	36	546	203	kappa-casein
MUSCRYGD	599	7	525	67	gamma-D-crystallin
MUSCTNCA	703	43	486	174	troponin C
MUSGBPA	478	19	408	51	galactoside binding
MUSGLOBZ	556	22	429	105	zeta-globin
MUSHIS3A	595	21	411	163	histone H3
MUSLACPI	844	41	675	128	placental lactogen
MUSMK2P	728	43	423	262	retinoic responsive
MUSNGF7S	830	15	771	44	nerve growth factor
Plant sequences					
BNANAP	718	42	537	139	napin
PEAABN1M	609	17	393	199	albumin 1
PHVCHM	1132	33	987	112	chitinase
PHVLBA	511	44	441	26	leghemoglobin
SOYCIIPI	425	30	312	83	CII protease inhibitor
SOYHSP176	718	96	465	157	low MW heat shock protein
TAHI02	626	62	411	153	histone H3
TOMRBCSD	778	73	546	159	RuBP carboxylase small
Amphibian sequences					
XELGSCHB	1069	114	732	223	goosecoid homeobox
XELHISH1	1180	277	822	81	histone H1 maternal
XELIGFIA	941	279	462	200	insulin-like factor
XELLBL	796	16	408	372	lactose binding
XELPCNA	1018	27	786	205	proliferating antigen
XELPYLA	411	64	195	152	PYLa precursor
XELRIGA	499	26	438	35	insulinoma
XELSRBP	892	40	594	258	retinol binding protein

Statistical significance was tested for the biases observed in calculated folding free energy between native and randomized sequences. The statistical significance of the differences in free energy was measured in standard deviation units, termed the segment score from Le and Maizel (10). Large sets of randomized mRNA folding free energies were found to be normally distributed. Standard hypothesis tests were employed using statistical analysis software. All thermodynamic energies are free energies expressed as kcal/mol. A greater negative free energy indicates that a more stable folding configuration is possible.

RESULTS

Fifty-one mRNA sequences were selected from a variety of plant, animal and bacteria sequences in GenBank (Table 1). These sequences and their respective randomized sequences were *in silico* folded using Zuker's MFOLD program. To ensure consistency of the folding algorithm with the selected set of mRNA sequences, the calculated native folding free energies were plotted as a function of sequence length in Figure 1. As expected, folding free energies become more negative for increasing sequence length since more bonding interactions are possible in longer molecules. A linear regression equation computed from data in Figure 1 gives a slope of –0.21 (kcal/mol per nucleotide) with an R-squared value of 0.58. This indicates that sequence length accounts for over half of the free energy of each mRNA. A further check of the data set examined normalized folding free energies (free energy divided by sequence length) as a function of C+G percent content in Figure 2. As expected the folding free energy per base becomes more negative as the C+G content of the mRNA increases due to the greater interaction energy between C-G pairs compared with A-U pairs. A linear regression equation computed from data in Figure 2 gives a slope of –0.005 (kcal/mol/base per %G+C) with a 0.66 regression coefficient. Thermodynamic data from others indicates that an A-U pair contributes –0.9 kcal/mol while a C-G pair contributes –2.9 kcal/mol (11).

The native folding free energies and the mean of free energies from sets of 10 randomized sequences are shown in Table 2 labeled as 'whole-random'. Of the 51 mRNAs examined, 40 (78%) are more negative than the mean of the whole-randomized set. Native mRNA sequences then are generally more stable than the corresponding whole-randomized sequences. A segment score calculated from the standard deviation of each randomized set is listed in Table 2. These values are the number of standard deviations the means of the randomized set are away from the native free energy. The average segment score for the whole-randomized set is –1.23 with a 95% confidence interval of 0.45, indicating a significant bias. Four mRNAs have segment scores greater than –4, while 13 (25% of the 51 mRNAs) have scores greater then –2. These cases indicate significantly greater folding stability of the native mRNA compared to randomized sequences.

Of the set of mRNAs with positive segment scores, none has a score greater than +2. Included in this set of mRNAs, less stable than expected are tomrbcsd (Rubisco) and xelhish1 (a histone). Interestingly, these classes of genes are suspected to be at least partially regulated by post-transcriptional events. Rubisco small subunit has been demonstrated to be post-transcriptionally regulated (12). The cell-cycle regulation of histone mRNA stability has also been demonstrated (13), so it is not unreasonable to suspect that the folding stability for these mRNAs may be different from other mRNAs not so regulated. In addition, equal

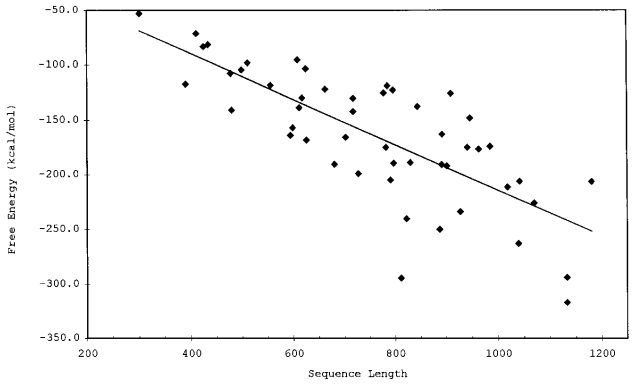

Figure 1. mRNA folding free energy dependence on sequence length. Data points are calculated free energy values of mRNA sequences listed in Table 1. Fitted regression line is shown.

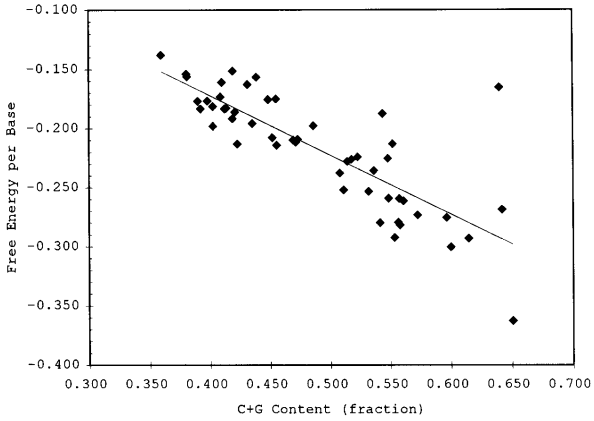

Figure 2. Sequence length normalized folding free energy dependence on G+C content. Units for free energy are kcal/mol/base. Fitted regression line is shown.

G and C contents in histone genes were hypothesized by Huynen to indicate selection pressures on mRNA secondary structure (14). The free energy of histone mRNA secondary structures was noted in Huynen's study to be only slightly lower than expected on the basis of nucleotide frequencies.

To determine if the above observed bias toward more stable mRNAs resides in the coding region or the flanking untranslated regions of the mRNAs, native sequences were randomized only within the coding region, yielding 'CDS-random' sequences. These sequences contain identical 5′ and 3′ UTRs of the respective native mRNA. Again the native mRNA sequences are usually more negative than the corresponding CDS-random sequences (Table 2). The average segment score is –0.87 with a 95% confidence interval of 0.48. Twelve (24% of the 51 mRNAs) have scores greater than –2. This demonstrates there is also a significant difference in folding free energies between native and partially randomized mRNA sequences. Of the 51 mRNAs examined, 37 or 73% are more negative than the CDS-randomized sequences. The 51 mRNAs possessed a total 5′ UTR length of 3014 nt, a total coding length of 27 306 nt and a total 3′ UTR length of 8533 nt. Therefore the coding regions comprise 70% of the total mRNA nucleotides, yet randomization of the coding region does not substantially alter the number of mRNAs observed with a bias toward more negative folding free energies. The significance level of this bias is also only slightly reduced by CDS-randomization compared to whole-sequence randomization. In only nine cases the sign of the segment

score changed. Six mRNA scores changed from minus to positive (indicating the native mRNA changed to being less stable than the randomized set), while three changed from positive to negative. In almost all of these changes, the segment scores were close to zero. Of the group with more negative folding free energies, 23 or 45% are significantly more negative by one standard deviation, and 13 or 25% are more negative by two standard deviations.

The above randomization procedure was modified to shuffle the codons while preserving the native amino acid composition and UTR sequences. These 'codon-shuffled' mRNAs again are generally less stable than the respective native mRNA sequence. Thirty-two of the 51 mRNAs (63%) have negative segment scores, with 13 (or 25%) being greater than –1. The mean of the segment scores is lower than the other two randomization procedures, yet the 95% confidence interval (0.47) still does not include zero.

To determine if the observed bias toward more stable mRNAs resides in the choice of codons within the coding sequence, a fourth randomization procedure was performed. Native sequences were randomized only by codon choice within the coding region, yielding 'codon-random' sequences with unmodified base composition. These sequences contained identical 5′ and 3′ UTRs of the respective native mRNA and coded for the same polypeptide. Again the native mRNA sequences tend to fold more negative in free energy than the corresponding codon-random sequences (Table 3). Since only a relatively small number of bases will change under this randomization procedure, the resulting sequences in the randomized set will be similar. As a consequence the folding free energies are very close in each set of randomized sequences, resulting in a very low standard deviation. This results in a very large segment score for several mRNAs, so instead the percent difference from the native free energy becomes a more appropriate measure of randomization effects. The mean of the percent difference of native from codon-random free energies is –2.2%, with a 95% confidence interval from –4.0 to –0.4%. The confidence interval does not include zero, demonstrating a significant difference in folding free energies between native and codon-randomized mRNA sequences. Of the 51 mRNAs examined, 35 or 69% are more negative than their codon-randomized mRNAs.

To investigate the effect of the choice of codons within the coding sequence, a fifth randomization procedure was performed. Native sequences were randomized by an equal selection of codons within the coding region, yielding 'codon-flat' sequences. These sequences contained identical 5′ and 3′ UTRs of the respective native mRNA and coded for the same polypeptide as with the codon-random sets. Since the choice of codons was unbiased, the resulting base composition was usually different from the native sequences. Again the native mRNA sequences tend to fold more negative in free energy than the corresponding codon-flat sequences (Table 3). The mean of the percent difference of native from codon-flat free energies is –6.6%, with a 95% confidence interval from –10.0 to –3.2%. The confidence interval does not include zero, demonstrating a significant difference in folding free energies between native and codon-flat mRNA sequences. Of the 51 mRNAs examined, 37 or 73% are more negative than their codon-flat mRNAs.

Much data suggest that UTRs of mRNA contain important RNA secondary structures. A great deal of evidence shows that the 5′ UTR is critically involved in regulating translation initiation (15). Alterations in translation regulation not only directly affect the amount of a protein that is finally synthesized, but can also significantly modify the stability characteristics of the message, and therefore modify protein levels by this mechanism as well. A non-viral example of a regulatory element in the 5′ UTR that has been well studied and has been shown to control message translation is the iron response element (IRE) that is found in ferritin mRNA (16). This sequence provides a portion of a stem–loop structure to which an iron regulatory protein (IRP) can bind (17). There are also many examples of elements in the 3′ UTR part of the message that bind to *trans*-acting proteins to control mRNA turnover rates (18). To detect the general presence of mRNA structures in the UTRs, a final randomization procedure was tried. Native sequences were randomized only within the UTRs, yielding 'UTR-random' sequences. Again the native mRNA sequences are usually more negative than the corresponding UTR-random sequences (Table 2). The average segment score is –0.50 with a 95% confidence interval of 0.39. Nine of the mRNAs have segment scores greater than –2. Although this randomization procedure produces lower segment scores than the CDS-random case, it is based on a smaller region for randomization. Since the UTRs comprise only 30% of the total mRNA nucleotides, the observed bias toward more negative folding free energies demonstrates there is still a significant difference in folding free energies between native and randomized mRNA sequences.

DISCUSSION

The biases in calculated mRNA folding free energies observed in Table 2 and 3 are small yet significant. For a 400 nt mRNA that is 50% basepaired and 50% G+C, a 5% increase in folding free energy could be caused by changes in only 7–20 bp. Since the CDS comprises ~70% of the mRNA examined, more of the bias is due to amino acid sequence and codon choice than due to the UTR sequences alone. If the amino acid sequence is constrained by considerations of protein function, then the bias is most likely due to subtle arrangements in codon choice. The effect observed here may be due to a selective advantage for mRNAs to be more basepaired, perhaps to resist degradation or modification. If indeed RNA was the original genetic material as suggested by the research of Joyce (19), then the genetic code may be arranged in such a manner as to encourage intermolecular bonding for single-stranded RNA. This may be of advantage for protoorganisms with genetic information stored in the more labile RNA compared to DNA.

The fact that these biases are observed from an empirical molecular calculation also suggests that local secondary structures are the causative agents. Most algorithms for predicting RNA secondary structure from base sequence are based on a nearest neighbor model of interaction (20,21). Experimental evidence indicates short-ranged stacking and hydrogen bonding are important determinants of RNA stability while hydrophobic bonding is of lesser importance (21). Numerous algorithms have been developed to predict RNA secondary structure by minimizing the configurational free energy. The quality of these predictions depends upon: (i) the accuracy of the thermodynamic data which describe the free energies of various secondary structural features; (ii) the folding rules that an algorithm uses to find the lowest free energy structure; and (iii) the degree to which environmental conditions stabilize alternate structures of equivalent or higher energy. Similar results were obtained using the older FOLD program, indicating the conclusions from this work are not sensitive to the algorithm nor energy data sets used.

Table 2. Folding free energies and segment scores for mRNAs

Name	Native	Whole-random	Segment score	CDS-random	Segment score	Codon-shuffled	Segment score	UTR-random	Segment score
Drocskb	–148.0	–144.3	–0.51	–149.5	0.21	–147.6	–0.09	–150.3	0.64
Drometo	–52.8	–49.2	–0.77	–55.0	0.43	–56.3	0.77	–50.1	–0.72
Drosist	–175.0	–184.5	1.24	–185.4	1.82	–182.8	1.19	–176.4	0.32
Drotu4a	–103.3	–118.4	1.79	–116.6	4.02	–115.3	1.38	–109.7	1.52
Droubxdr	–121.6	–95.1	–4.94	–104.2	–3.93	–107.1	–3.45	–121.4	–0.05
Drovmp	–81.3	–93.5	1.40	–93.1	2.22	–93.7	2.43	–85.6	2.18
Ecoadd	–263.2	–245.6	–4.86	–256.1	–0.93	–258.8	–0.66	–259.1	–3.45
Ecoalka	–250.0	–246.6	–0.35	–245.3	–0.45	–255.0	0.58	–254.4	1.47
Ecocma	–191.8	–154.3	–4.03	–159.1	–4.70	–161.2	–8.38	–189.8	–0.63
Ecodapa	–233.9	–218.6	–1.63	–225.4	–1.72	–226.6	–0.74	–231.9	–0.76
Humalr	–316.8	–290.1	–2.68	–294.1	–2.60	–298.5	–2.19	–311.2	–1.12
Humcal	–204.9	–193.1	–1.54	–194.9	–1.94	–196.1	–1.61	–207.0	0.35
Humcalci	–190.4	–175.0	–2.61	–178.2	–2.70	–187.7	–0.35	–179.9	–2.15
Humgrp5e	–189.4	–180.9	–0.76	–181.8	–1.70	–188.6	–0.15	–186.7	–0.44
Humgst	–125.4	–129.3	0.49	–126.9	0.21	–126.3	0.30	–128.5	0.74
Humhembp	–240.3	–218.3	–3.18	–216.6	–3.92	–220.3	–3.13	–228.8	–1.79
Humhis4	–117.1	–113.2	–0.90	–117.0	–0.03	–114.4	–0.41	–115.9	–0.42
Humhpbs	–294.1	–272.7	–2.04	–281.9	–1.59	–295.3	0.21	–278.3	–2.46
Humifnab	–206.0	–173.5	–4.90	–181.6	–4.28	–188.0	–2.68	–194.1	–2.21
Humifnac	–176.5	–153.7	–2.75	–163.3	–1.38	–169.9	–0.85	–168.9	–1.49
Humifnaf	–174.2	–145.7	–2.75	–164.5	–1.67	–171.5	–0.42	–165.5	–2.64
Humifnah	–174.0	–146.4	–3.41	–153.3	–2.71	–162.4	–1.28	–167.5	–1.67
Huml12a	–138.3	–128.4	–1.62	–132.0	–1.13	–137.1	–0.15	–137.7	–0.12
Humogc	–190.6	–181.2	–1.09	–191.7	0.20	–194.6	0.77	–170.7	–2.43
Hump1bx	–140.7	–141.3	0.10	–141.6	0.11	–140.5	–0.06	–139.7	–0.13
Mmu03711	–129.5	–126.9	–0.48	–130.2	0.09	–128.4	–0.16	–134.6	1.28
Muscask	–118.7	–101.3	–1.90	–107.2	–2.17	–111.3	–0.89	–109.8	–2.30
Muscrygd	–156.7	–150.2	–0.72	–148.9	–1.03	–144.8	–3.14	–154.8	–0.76
Musctnca	–165.6	–181.0	1.66	–176.5	1.36	–169.0	0.33	–161.1	–0.88
Musgbpa	–107.6	–105.2	–0.54	–106.1	–0.33	–107.0	–0.10	–105.5	–0.72
Musglobz	–118.3	–114.8	–0.44	–124.3	0.85	–120.1	0.27	–122.5	1.86
Mushis3a	–164.0	–155.5	–0.87	–161.5	–0.27	–161.2	–0.32	–164.9	0.14
Muslacpi	–137.3	–144.4	0.95	–144.4	1.16	–146.4	1.06	–141.8	0.67
Musmk2p	–199.1	–188.3	–1.90	–187.9	–2.31	–194.4	–0.93	–190.8	–2.03
Musngf7s	–188.9	–185.7	–0.34	–179.2	–0.91	–177.1	–1.74	–186.1	–0.94
Bnanap	–141.9	–126.8	–1.78	–141.0	–0.14	–142.7	0.15	–138.0	–1.44
Peaabn1m	–95.2	–93.4	–0.28	–94.7	–0.08	–98.4	0.69	–92.5	–0.64
Phvchm	–293.7	–274.2	–2.29	–279.7	–1.44	–287.4	–0.62	–291.8	–0.72
Phvlba	–97.9	–86.0	–1.90	–86.8	–1.32	–94.2	–0.56	–95.5	–2.23
Soyciipi	–83.2	–76.4	–0.71	–83.4	0.02	–85.2	0.44	–76.2	–1.70
Soyhsp176	–130.1	–123.8	–0.95	–127.5	–0.49	–131.3	0.29	–122.2	–1.45
Tahi02	–168.2	–171.9	0.39	–166.7	–0.35	–168.5	0.03	–177.6	2.40
Tomrbcsd	–125.1	–126.3	0.17	–136.1	1.75	–133.7	1.01	–129.0	1.48
Xelgschb	–225.9	–211.4	–1.87	–224.6	–0.12	–228.0	0.27	–222.7	–0.60
Xelhish1	–206.3	–207.1	0.09	–194.3	–1.21	–218.4	0.84	–212.6	0.89
Xeligfia	–175.1	–152.5	–3.05	–163.1	–2.01	–162.1	–2.31	–162.6	–1.68
Xellbl	–122.6	–119.4	–0.37	–121.9	–0.09	–121.5	–0.26	–119.5	–0.80
Xelpcna	–211.3	–195.3	–1.39	–191.6	–2.59	–195.1	–2.44	–208.1	–0.50
Xelpyla	–71.2	–65.2	–0.96	–60.3	–3.28	–63.4	–1.61	–75.7	1.10
Xelriga	–104.3	–104.5	0.03	–101.3	–0.42	–98.0	–0.90	–107.5	1.40
Xelsrbp	–163.0	–155.9	–1.03	–157.2	–1.11	–153.9	–2.59	–163.8	0.18
Average			–1.23		–0.87		–0.63		–0.50
Confidence interval (95%)			0.45		0.48		0.47		0.39

Table 3. Folding free energies and percent difference for mRNAs

Name	Native	Codon-random	Segment score	Difference (%)	Codon-flat	Segment score	Difference (%)
Drocskb	–148.0	–146.7	–0.55	0.9	–141.8	–0.61	4.2
Drometo	–52.8	–53.8	0.39	–1.9	–49.4	–1.12	6.4
Drosist	–175.0	–174.1	–0.24	0.5	–148.8	–3.18	15.0
Drotu4a	–103.3	–122.8	55.01	–18.8	–140.7	3.45	–36.2
Droubxdr	–121.6	–106.9	–1.82	12.1	–97.0	–3.45	20.3
Drovmp	–81.3	–89.4	5.21	–10.0	–91.5	1.67	–12.6
Ecoadd	–263.2	–260.0	–0.27	1.2	–243.8	–1.79	7.4
Ecoalka	–250.0	–246.2	–1.11	1.5	–223.1	–2.29	10.7
Ecocma	–191.8	–159.8	–2.83	16.7	–182.0	–1.08	5.1
Ecodapa	–233.9	–221.9	–3.39	5.1	–198.2	–3.64	15.3
Humalr	–316.8	–292.5	–16.40	7.7	–253.6	–5.79	19.9
Humcal	–204.9	–196.8	–1.36	4.0	–176.9	–3.68	13.7
Humcalci	–190.4	–178.0	–1.57	6.5	–158.6	–8.85	16.7
Humgrp5e	–189.4	–183.8	–2.64	3.0	–165.5	–3.10	12.6
Humgst	–125.4	–128.7	0.25	–2.6	–138.3	1.69	–10.3
Humhembp	–240.3	–217.9	–12.70	9.3	–198.3	–5.20	17.5
Humhis4	–117.1	–120.0	0.38	–2.4	–91.6	–2.87	21.8
Humhpbs	–294.1	–287.1	–1.55	2.4	–246.9	–5.26	16.1
Humifnab	–206.0	–181.3	–3.30	12.0	–171.3	–3.99	16.9
Humifnac	–176.5	–161.1	–4.12	8.8	–160.4	–2.42	9.1
Humifnaf	–174.2	–159.2	–19.35	8.6	–163.0	–1.37	6.5
Humifnah	–174.0	–156.1	–0.93	10.3	–149.4	–2.70	14.1
Huml12a	–138.3	–133.6	–3.32	3.4	–127.4	–1.65	7.9
Humogc	–190.6	–199.3	1.21	–4.6	–196.8	1.52	–3.3
Hump1bx	–140.7	–143.1	0.65	–1.7	–124.7	–2.66	11.4
Mmu03711	–129.5	–130.2	0.71	–0.5	–121.9	–0.85	5.9
Muscask	–118.7	–110.3	–2.17	7.1	–119.6	0.10	–0.7
Muscrygd	–156.7	–147.9	–0.85	5.6	–121.8	–5.94	22.2
Musctnca	–165.6	–165.8	0.04	–0.1	–143.1	–3.62	13.6
Musgbpa	–107.6	–114.3	1.39	–6.2	–93.8	–1.62	12.8
Musglobz	–118.3	–114.8	–0.51	3.0	–103.0	–2.55	13.0
Mushis3a	–164.0	–165.2	0.86	–0.7	–127.3	–8.01	22.4
Muslacpi	–137.3	–135.9	–0.41	1.0	–149.2	2.07	–8.6
Musmk2p	–199.1	–195.3	–1.34	1.9	–171.2	–2.74	14.0
Musngf7s	–188.9	–173.3	–11.65	8.3	–161.9	–4.86	14.3
Bnanap	–141.9	–138.6	–1.10	2.4	–139.5	–0.64	1.7
Peaabn1m	–95.2	–101.8	92.63	–6.9	–119.1	3.87	–25.1
Phvchm	–293.7	–284.4	–5.98	3.2	–257.6	–4.44	12.3
Phvlba	–97.9	–101.1	1.94	–3.2	–100.4	0.41	–2.6
Soyciipi	–83.2	–76.7	–1.53	7.8	–87.3	0.74	–5.0
Soyhsp176	–130.1	–135.7	0.95	–4.3	–128.3	–0.42	1.4
Tahi02	–168.2	–168.0	–0.24	0.1	–134.8	–3.32	19.8
Tomrbcsd	–125.1	–136.0	1.33	–8.7	–130.5	0.84	–4.3
Xelgschb	–225.9	–213.0	–2.28	5.7	–213.3	–1.51	5.6
Xelhish1	–206.3	–219.6	4.80	–6.4	–215.9	0.90	–4.6
Xeligfia	–175.1	–167.9	–0.96	4.1	–176.4	0.21	–0.8
Xellbl	–122.6	–118.2	–3.46	3.6	–127.5	1.34	–4.0
Xelpcna	–211.3	–194.6	–4.16	7.9	–192.4	–1.58	8.9
Xelpyla	–71.2	–63.5	–2.27	10.8	–74.6	0.63	–4.8
Xelriga	–104.3	–102.3	–1.41	1.9	–90.9	–2.06	12.9
Xelsrbp	–163.0	–157.0	–0.51	3.7	–158.3	–0.73	2.9
Average			0.97	2.2		–1.81	6.5
Confidence interval (95%)				1.8			3.4

Free energy minimizing algorithms such as Zuker's MFOLD program output a family of structures that have the same or nearly the same free energy. This study has compared the optimal free energy of folding with a reference set of optimal free energies obtained by sequence randomization. Thus what are being compared are the locations of the minima in the configurational energy profiles for folding. Different regions of mRNA sequence were randomized, including codon choice, resulting in destabilization of the folding free energy. The under-representation of 5′-CG-3′ doublets in most native sequences would be changed upon such randomization. The 5′-CG-3′ doublet contributes ~2 kcal/mol in helix propagation, whereas the 5′-GC-3′ and 5′-GG-3′ doublet each contribute ~3 kcal/mol (from the GCG energy file for MFOLD). Even though randomization often increases the number of 5′-CG-3′ doublets compared with 5′-GC-3′ doublets, there is no correlation seen between 5′-CG-3′ content and free energy within the randomized sets. Pearson coefficients of free energy versus CG-doublet content are positive and negative, with magnitudes mostly below 0.25 (data not shown). This suggests that local secondary structure interactions are causing the observed bias in folding free energies, not changes in CG content.

This bias in mRNA folding free energy calculations has potential applications for gene finding algorithms. It is not known if open reading frames (ORFs) that are not transcribed show the bias observed in Table 2. It would be difficult to ensure that an ORF was not a region of some mRNA, due to the limited (and sometimes incorrect) information contained in sequence annotation. Although the folding calculations are computationally intensive, the results may be useful to improve the predictive accuracy of gene finding programs. The same level of bias for the same set of mRNAs was observed using the faster (but older) FOLD program (data not shown).

The thermodynamic treatment in this work concerns not the actual mRNA structures but the depth of the free energy potential well that a mRNA molecule could fold due to its sequence. The sequence of mRNAs has been found in this study to generally give rise to more stable secondary structures than expected by chance. Codon choice has been shown to contribute to this observed bias. Although the average bias is statistically significant for all of the randomization procedures used, the average segment scores are not strong in the sense of being greater than –3.0 or more. Yet individually, several sequences were found with very strong segment scores greater than –4.0. The conclusions drawn, however, must be tempered by the fact that the set of genes examined was small considering the 51 sequences were grouped into categories such as invertebrate, human, etc., leaving few sequences per category to draw major conclusions. These results could be useful to develop a classification of mRNAs based on whether mRNAs are more or less stable compared to their randomized sequences. Gene regulation mechanisms or gene product types may be related to the mRNA folding class based on one of the five randomization procedures examined here.

CONCLUSION

A survey of 51 mRNA sequences reveals a bias in the coding and UTRs that allows for greater negative folding free energies than predicted by sequence length or nucleotide base content. A free energy reference state is taken to be a large enough set of randomized mRNA sequences. Randomization can be implemented over the whole sequence or over sections such as the CDS, UTR or codon choice. Randomization of most regions of the mRNA sequences display lower folding stability as measured by calculated free energy values. Randomization of codon choice while still preserving original base composition also results in less stable mRNAs. This suggests that a bias in the selection of codons favors mRNA structures which contribute to folding stability.

ACKNOWLEDGEMENTS

I thank Jonathan Arnold (University of Georgia) for helpful discussions concerning this work. This work was supported (or partially supported) by NIH grant GM08247, by a Research Centers in Minority Institutions award, G12RR03062, from the Division of Research Resources, National Institutes of Health, and NSF CREST Center for Theoretical Studies of Physical Systems (CTSPS) Cooperative Agreement #HRD-9632844.

REFERENCES

1 de Smit,M.H. and van Duin,J. (1990) *Prog. Nucleic Acid Res. Mol. Biol.*, **38**, 1–35.
2 Jacobson,A.B., Arora,R., Zuker,M., Priano,C., Lin,C.H. and Mills,D.R. (1998) *J. Mol. Biol.*, **274**, 589–600.
3 Love,H.D.,Jr, Allen-Nash,A., Zhao,Q. and Bannon,G.A. (1988) *Mol. Cell. Biol.*, **8**, 427–432.
4 Turner,P.C. and Woodland,H.R. (1982) *Nucleic Acids Res.*, **10**, 3760–3780.
5 Wells,D.E. (1986) *Nucleic Acids Res.*, **14** (suppl.), r119–r150.
6 Lloyd,A.T. and Sharp,P.M. (1992) *Nucleic Acids Res.*, **20**, 5289–5295.
7 Bains,W. (1987) *J. Mol. Biol.*, **197**, 379–388.
8 Devereux,J., Haeberli,P. and Smithies,O. (1984) *Nucleic Acids Res.*, **12**, 387–395.
9 Zuker,M. and Stiegler,P. (1981) *Nucleic Acids Res.*, **9**, 133–148.
10 Le,S.-Y. and Maizel,J.V.,Jr (1989) *J. Theor. Biol.*, **138**, 495–510.
11 Freier,S., Kierzek,R., Jaeger,J., Sugimoto,M., Caruthers,M., Neilson,T. and Turner,D. (1986) *Proc. Natl Acad. Sci. USA*, **83**, 9373–9377.
12 Berry,J.O., Nikolan,B.J., Carr,J.P. and Klessig,D.F. (1985) *Mol. Cell. Biol.*, **5**, 2238–2246.
13 Graves,R.A., Pandey,N.B., Chodchoy,N. and Marzluff,W.F. (1987) *Cell*, **4**, 615–626.
14 Huynen,M.A., Konings,D.A.M. and Hogeweg,P. (1992) *J. Mol. Evol.*, **34**, 280–291.
15 Ross,J. (1995) *Microbiol. Rev.*, **59**, 423–450.
16 Klausner,R.D. and Harford,J.B. (1989) *Science*, **246**, 870–872.
17 Henderson,B.R. and Kuhn,L.C. (1995) *J. Biol. Chem.*, **270**, 20509–20515.
18 Hake,L.E. and Richer,J.D. (1997) *Biochim. Biophys. Acta*, **1332**, M31–M38.
19 Joyce,G. (1989) *Nature*, **338**, 217–224.
20 Walter,A.E., Turner,D.H., Kim,J., Lyttle,M.H., Müller,P., Mathews,D.H. and Zuker,M. (1994) *Proc. Natl Acad. Sci. USA*, **91**, 9218–9222.
21 Mathiews,D.H., Andre,T.C., Kim,J., Turner,D.H. and Zuker,M. (1998) In Leontis,N.B. and SantaLucia,J.,Jr (eds), *Molecular Modeling of Nucleic Acids*. American Chemical Society Symposium Series 682, Washington, DC, pp. 246–257.